Syst. Biol. 59(4):465–476, 2010

DOI:10.1093/sysbio/syq026
Advance Access publication on May 28, 2010

Sampling Trees from Evolutionary Models

KLAAS HARTMANN[1], DENNIS WONG[2], AND TANJA STADLER[3,*]

[1]*Tasmanian Aquaculture and Fisheries Institute, University of Tasmania, Hobart, Australia;* [2]*Faculty of Graduate Studies, Dalhousie University, Halifax, Canada; and* [3]*Institute of Integrative Biology, Eidgenössische Technische Hochschule Zürich, Universitätsstrasse 16, 8092 Zürich, Switzerland;*
[*]*Correspondence to be sent to: Institute of Integrative Biology, ETH Zürich, Universitätsstrasse 16, 8092 Zürich, Switzerland; E-mail: tanja.stadler@env.ethz.ch.*

Received 17 March 2008; reviews returned 22 May 2008; accepted 17 February 2010
Associate Editor: Tim Collins

Abstract.—A wide range of evolutionary models for species-level (and higher) diversification have been developed. These models can be used to test evolutionary hypotheses and provide comparisons with phylogenetic trees constructed from real data. To carry out these tests and comparisons, it is often necessary to sample, or simulate, trees from the evolutionary models. Sampling trees from these models is more complicated than it may appear at first glance, necessitating careful consideration and mathematical rigor. Seemingly straightforward sampling methods may produce trees that have systematically biased shapes or branch lengths. This is particularly problematic as there is no simple method for determining whether the sampled trees are appropriate. In this paper, we show why a commonly used simple sampling approach (SSA)—simulating trees forward in time until n species are first reached—should only be applied to the simplest pure birth model, the Yule model. We provide an alternative general sampling approach (GSA) that can be applied to most other models. Furthermore, we introduce the constant-rate birth–death model sampling approach, which samples trees very efficiently from a widely used class of models. We explore the bias produced by SSA and identify situations in which this bias is particularly pronounced. We show that using SSA can lead to erroneous conclusions: When using the inappropriate SSA, the variance of a gradually evolving trait does not correlate with the age of the tree; when the correct GSA is used, the trait variance correlates with tree age. The algorithms presented here are available in the Perl Bio::Phylo package, as a stand-alone program TreeSample, and in the R TreeSim package. [Algorithms; distribution; evolutionary models; phylogenetic trees; sampling; simulating.]

Evolutionary models have been developed for many reasons. One of their main uses has been to try to explain the evolution of biological diversity for organisms. Studies in this field fit and compare a developed model with a data set (a record of fossil presence through time or a phylogeny). Comparing models of evolution with a data set is an important part of hypothesis testing and an integral part of the scientific method (for example, see Sepkoski 1982; Bininda-Edmonds et al. 2007; and for a review, see Mooers and Heard 1997; Mooers et al. 2007). In this paper, we investigate the issue of comparing an evolutionary model with a reconstructed phylogeny of present-day species.

A reconstructed phylogeny is a tree of age t, containing n extant species. This reconstructed phylogeny is compared with model trees with the same age (t) and number of extant species (n). The properties of the trees predicted by most evolutionary models cannot be described analytically, hence it is necessary to simulate the trees. One starts at time t in the past and evolves trees under the model up to the present day. As n species are observed at the present day, only simulations with n present-day species shall be considered and compared with the reconstructed phylogeny. The time t in the past is an inferred uncertain value in the phylogeny, and therefore it is often preferable to only condition the simulations on the number of extant species (n) and not the age of the tree (t).

There are numerous ways to produce trees with a given number of species from an evolutionary model. The most widely used simple sampling approach (SSA) starts with a single species and evolves a tree until n species are reached. The simulation is stopped with the next speciation or extinction event.

We show that the SSA is appropriate for the widely used Yule and coalescent models, however there are fundamental problems applying these approaches to other evolutionary models. Most obviously, later periods with n species are disregarded and pendant edges (i.e., edges adjacent to the leaves) are too long due to stopping after the interval of observing n species. Hence, the SSA produces trees with systematically biased shapes and/or branch lengths. Most apparently, 1) stopping at the next speciation or extinction event is extreme and suggests a bias in the pendant edge lengths (i.e., edges adjacent to the leaves) and 2) for models that explicitly include extinction, the SSA disregards possible later periods where n species are extant.

In this paper, alternative general sampling approaches (GSAs) are provided that are theoretically sound and easy to apply even for complex models. Furthermore, we provide a very fast and efficient approach for sampling under a constant-rate birth–death process model, the birth–death sampling approach (BDSA). We investigate the importance of using our new sampling approaches over the established methods for the constant-rate birth–death models. This is achieved by comparing samples of trees produced by the different sampling approaches. The SSA induces a strong bias in the age of a tree (SSA trees are too young) and a less pronounced bias in the relative timing of the speciation events (the direction of the bias depends on the ratio of extinction rate and speciation rate). Furthermore, the

SSA induces a negligible bias in the tree shape distribution for incomplete taxon sampling. We identify attributes of other models that will result in the SSA producing more biased samples.

We conclude the paper by showing that using inappropriate sampling approaches can lead to erroneous conclusions about evolutionary mechanisms. This is demonstrated using a case study first considered by Purvis (2004), where the correlation between the age of a tree and the variance of an evolving trait is examined. Under a punctuational trait evolution model, there is no correlation. However, under a gradual trait evolution model, trees sampled using the SSA show no correlation and trees sampled using the correct GSA show significant correlation. This shows that using exact sampling methods not only is of theoretical interest but also can have major effects on conclusions in data analysis: A study based on SSA would suggest that data with no correlation between trait variance and tree age are compatible with a gradual trait evolution model, whereas a study based on the correct GSA would suggest otherwise.

The sampling methods we present are not the fastest or most sophisticated; however, in our opinion they are the easiest to implement and applicable to the broadest possible range of models. Most of our algorithms are implemented in the Perl Bio::Phylo package as well as the R TreeSim package, where they can easily be applied to any suitable evolutionary model. For those users unfamiliar with Perl or R, we have also made them available using a stand-alone GUI TreeSample. The Perl tools and TreeSample are freely available from Hartmann (2010). The R package TreeSim can be downloaded from Stadler (2010). Lastly, we note that although we present our work in the context of evolutionary models of species diversification, our methods can be applied to other scenarios where birth–death processes are modeled, for example, gene trees (Karev et al. 2003; Hahn et al. 2005; Oakley et al. 2006) or transmission of infectious diseases (Tanaka et al. 2006).

Sampling Methods

Throughout this paper, the aim is to produce a sample from the tree probability distribution induced by an evolutionary model. The first problem is that this tree probability distribution is ill defined for most evolutionary models. Under most models, trees evolve perpetually and trees of all ages are possible, hence the expected age of the tree (the time between the root and the tips) is infinite. To obtain a probability distribution, it is therefore necessary to condition on some aspect of the tree; the number of species and the age of the tree are arguably the 2 most common and useful choices.

Conditioning on the age of a tree, t, is appropriate if we wish to compare a model with trees of known age or want to test methods on simulated trees of a given age. It is relatively easy to sample trees of age t from an evolutionary model. The tree is simply evolved according to the model until it has reached the desired age. This process is repeated until a sufficient number of trees have been sampled.

Conditioning on the number of species, n, in a tree $\mathcal{T}$ may be commonly required. The age of a reconstructed tree may only be known with limited accuracy, however the number of species in the (reconstructed) tree is fixed. Consequently, it may be more appropriate to use samples from an evolutionary model with a fixed number of species (we also consider incomplete taxon sampling). Sampling from the tree distribution conditional on the number of species, $p(\mathcal{T}|n)$, is the basis of this paper.

Throughout this paper, we assume a uniform prior on the age of the tree as in Popovic (2004), Aldous and Popovic (2005), and Gernhard (2008). Consider a large number of simulation runs that begin at a uniformly distributed time before the present. Trees obtained by selecting only those simulations that have n species at the present are a sample from $p(\mathcal{T}|n)$. This is a convenient way of interpreting the distribution but is not a practical sampling approach as the simulation starting time is taken from an ill-defined distribution (between an infinite time in the past and the present). A given model (and its parameters) will induce a distribution on the age of the tree given its size. All our knowledge about the age of a tree is encapsulated in the model and the chosen parameter values; the uniform prior on the tree age represents the fact that we have no further knowledge about the tree age outside of these parameters.

Current Approaches

One SSA for sampling trees with n species has seen wide usage. With this approach, a tree is evolved under the model until it has n species. The length of the pendant edges is the time until the next event (speciation or extinction), which is disregarded. This approach produces trees conditional on the next event (speciation or extinction) occurring immediately after the end of the tree, which, we show here, is generally not the same distribution as $p(\mathcal{T}|n)$. It is difficult to justify this approach as it produces a sample of trees equivalent to what we would expect if all "real" trees were observed immediately prior to a speciation or an extinction event.

"PhyloGen" (Rambaut 2002) is a freely available tree sampler that has been used in a number of studies, for example, Shaw et al. (2003), Venditti et al. (2006), Weir (2006), and Hohl and Ragan (2007). It permits users to sample trees from constant-rate birth–death processes and episodic speciation models. These trees are conditioned on the age of the tree or the number of species, n. Conditioning on n in PhyloGen simply terminates a tree after it first reaches n species. Trees sampled with PhyloGen are younger than expected for our interpretation of $p(\mathcal{T}|n)$, and the pendant edges are shorter than expected—in fact, the species produced by the last speciation event have 0 length edges. If the last speciation event is removed (creating a tree with $n - 1$ species), sampling trees with PhyloGen is equivalent to SSA with

$n-1$ species when we have a model without extinction. Due to this similarity, throughout the remainder of this paper, we only consider SSA. There are 3 main possible problems with SSA and PhyloGen.

Problem 1. As we have noted, the pendant edges produced by SSA and PhyloGen have what appear to be extreme values. PhyloGen pendant edges seem to be too short, whereas those produced by SSA appear to be too long.

Problem 2. SSA and PhyloGen stop evolving the tree during (or just after) the first period of time where the tree has n species. For models with extinction, the number of species will fluctuate up and down so there may be many periods during which the tree has n leaves. For such models, SSA and PhyloGen will result in younger trees than expected.

Problem 3. A final concern with SSA and PhyloGen is that each model simulation run makes the same contribution to the final sample—one single tree. However, from our interpretation of $p(\mathcal{T}|n)$, the probability of observing a given simulation depends on the duration for which the simulated tree had n species—for example, if this duration is short, it is unlikely that the simulated tree will be observed although it has n species.

Pure-Birth Memoryless Models

We begin by considering pure-birth memoryless models—models that do not explicitly include extinction (pure birth) and where future evolution depends only on the number of extant species (memoryless). This class of models is of particular interest as an approach similar to SSA can be used to correctly sample phylogenetic trees from them. Furthermore, this class of models includes the most widely used speciation model—the Yule model (Yule 1924; Harding 1971)—and the most widely used null model in population genetics—the coalescent model (Kingman 1982a, 1982b, 1982c).

Under the Yule model, each species has the same probability of speciating per unit time and this speciation rate is constant over time. Consequently, the time between speciation events is exponentially distributed with parameter $m\lambda$, where m is the number of species that are extant and λ is the intrinsic rate of speciation. The coalescent model is derived from population genetics principles but is essentially the same as the Yule model with one exception—the time between coalescent events is exponentially distributed with parameter $\binom{m}{2}$ (in the following, we will use "speciation" for both speciation and coalescent events).

In this section, we show that although SSA is generally inappropriate for pure-birth memoryless models, it is actually a correct approach for the Yule model and the coalescent model. As these models are pure-birth models, there will only be one period during which n species exist, so Problem 2 does not apply. This leaves Problems 1 and 3 that we will show cancel each other out under the Yule model and the coalescent model. We speculate that the suitability of SSA for sampling from the most widely used null models has led to its application to other models for which it is unsuitable.

An important aspect of memoryless models is that evolution after the speciation event that created the nth species (s_n) is completely independent of the evolution that occurred up to that point. Consequently, it is possible to simulate trees from these models in 2 separate stages. First, using the model, a tree is simulated to the speciation event that created the nth species (denoted by s_n; see Fig. 1). A length λ is then added onto the pendant edges to produce the final tree. Due to the independence of these 2 processes, Problems 1 and 3 do not affect the simulation to s_n and are addressed entirely by an appropriate choice of λ. This raises the question from what probability density, $h(\lambda)$, the additional time λ should be sampled.

We begin by noting that any pure-birth memoryless model can be uniquely defined by the probability densities of the intervals between speciation events. We denote the time between the speciation event that created the nth and the $(n+1)$th species by σ_n (the time between s_n and s_{n+1}) and its probability density by $g_n(\sigma_n)$. Note that SSA makes the assumption that

$$h(\lambda) = g(\lambda).$$

This effectively produces a tree with n species conditional on the next speciation event occurring immediately—clearly not what was intended.

A seemingly better (but still generally incorrect) approach would be to simulate the tree until s_{n+1} and randomly terminate the tree between s_n and s_{n+1} (because all trees between these 2 events should be equally

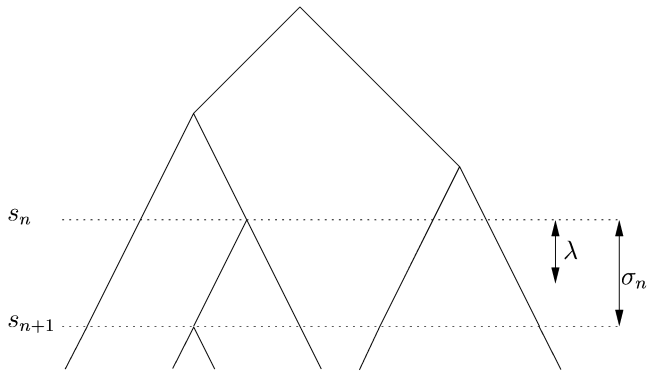

FIGURE 1. Some of the notation used throughout this paper is illustrated in this figure where $n = 5$. τ is the simulated tree until the point in time when the tree first has greater than n species. This point in time is the speciation event creating the $(n + 1)$th species—s_{n+1}. The duration for which a simulated tree has n species is denoted by σ_n, this is the time between the creation of the nth species (s_n) and the $(n + 1)$th species (s_{n+1}). The time for which an observed tree has n species is necessarily less than σ_n and is denoted by λ.

likely). This addresses Problem 1 and gives us

$$h(\lambda) = \int_{\sigma_n=\lambda}^{\sigma_n=\infty} h(\lambda|\sigma_n) g_n(\sigma_n) \mathrm{d}\sigma_n$$

$$= \int_{\sigma_n=\lambda}^{\sigma_n=\infty} \frac{g_n(\sigma_n)}{\sigma_n} \mathrm{d}\sigma_n.$$

However, this does not take into account the variable contribution to the $p(\mathcal{T}|n)$ that different values of σ_n should make (Problem 3).

From the definition of $p(\mathcal{T}|n)$, the contribution from a simulated tree with a given σ_n should be proportional to σ_n; therefore, the correct distribution from which to sample λ is

$$h(\lambda) \propto \int_{\sigma_n=\lambda}^{\sigma_n=\infty} \sigma_n h(\lambda|\sigma_n) g_n(\sigma_n) \mathrm{d}\sigma_n,$$

$$\propto \int_{\sigma_n=\lambda}^{\sigma_n=\infty} g_n(\sigma_n) \mathrm{d}\sigma_n. \quad (1)$$

Thus, the following will produce correct samples from $p(\mathcal{T}|n)$ for any pure-birth memoryless model:

Pure-birth memoryless sampling approach (PBMSA)

1. Simulate a tree terminating at s_n
2. Add a distance, λ, to the pendant edges using the correct $h(\lambda)$ from Equation 1
3. Repeat from Step 1 until all samples are obtained.

For SSA to be appropriate, we require $h(\lambda) = g_n(\lambda)$. Inspection of Equation 1 reveals that this requirement is met if $g_n(\sigma_n)$ is an exponential distribution. Furthermore, as the model is memoryless, the parameter may depend only on the number of species that are extant. These conditions are clearly satisfied by the Yule model, the coalescent model, and the related Moran (Moran 1958) and Hey models (Hey 1992).

Pure-birth memoryless sampling approach (PBMSA) is appropriate for any model where the time between speciation events depends only on the number of extant species though the Yule model and the coalescent model are the only widely used models that fit this category. PBMSA is inappropriate for models with explicit extinction events and models with a memory.

Explicit extinction events will result in a simulated tree that may have n species for several intervals—PBMSA would only sample from the first of these intervals resulting in a tree that is younger than expected.

Many models feature a memory, this may be in the form of hereditary speciation rates, for example, Heard (1996), or a dependence of speciation rates on the absolute age of a tree or a species, for example, Chan and Moore (1999). PBMSA cannot sample from such models as the evolution before and after s_n is not independent and different simulations to s_n should make different contributions to the final sample. If, for example, speciation rate is negatively related to species age (Agapow et al. 2004), then a tree of n old species should remain that size longer than a tree of n young species. Therefore, by the definition of $p(\mathcal{T}|n)$, a tree of n old species should give a greater contribution to that density than a tree of n young species. Consequently, it is necessary to take different numbers of samples from each of the evolutionary histories and PBMSA cannot be used.

A General Sampling Approach

We now introduce a general sampling method that works for a broad class of models that can include both speciation and extinction events. Our sampling approach simulates a tree, τ, until it is highly unlikely that the tree will return to n species. This will occur either when all species are extinct or when there has been sufficient speciation such that the number of extinctions required to return to n species is highly improbable.

The only restriction on the class of models from which our algorithm can sample is that we must be able to guarantee that each simulation "run" will eventually terminate. The efficiency of the algorithm depends on the time that is required until a simulation terminates. An example of a model to which this algorithm cannot be applied is one where the number of species perpetually fluctuates over a range including n.

Determining how unlikely a tree is to return to n species depends on the model. Throughout the remainder of this section, we assume that we can determine a critical number of species, n^*, from which it is unlikely that extinctions will bring the number of species back to n. A simulation therefore ends when the number of species reaches 0 or n^*. The value n^* can be obtained via simulations.

A simulation run will have k periods during which n species were extant, we denote the length of each of these periods by $\phi_i, i=1,\ldots,k$. As previously discussed, the probability of observing a simulated tree while it has n species is directly proportional to the duration for which n species existed: $\Phi = \sum_{i=1}^{k} \phi_i$. This will vary between simulations so each simulation should make a different contribution to the final sample—a simulated tree where n species existed for a short period of time should make a lower contribution to the sample than a simulated tree where n species existed for a longer period.

The question remains how to decide on the number of samples to take from a given simulated tree: this should be proportional to Φ. To take this into account, we introduce a sampling rate, r, such that we will take $r\Phi$ sampled trees from a given simulated tree. As we can only take whole samples of trees, for each simulated tree, $r\Phi$ will be randomly rounded: If $r\Phi$ is between integers k and $k+1$, it is rounded down with probability $r\Phi - k$ and up with probability $1 - (r\Phi - k)$. This ensures that the randomly rounded $r\Phi$ has an expected value of $r\Phi$.

If the sampling rate is too low, many simulations will be required for each sampled tree and the process will

be very inefficient. If it is too high, many sampled trees may be derived from a single simulated tree and these sampled trees will have a higher degree of correlation than expected for random samples. Ideally, r should be determined experimentally (by simulations) such that it is as high as possible while ensuring that few simulated trees produce more than a single sample. Like n^*, an appropriate value for r can be obtained from simulations.

Lastly, we introduce $S_i(\tau)$ as the set of trees that can be obtained by truncating a simulated tree during the ith interval during which it had n species. Combining these element, we have the following sampling approach:

General sampling approach

1. Determine a suitable sampling rate, r
2. Simulate a tree, τ, until n^* species or extinction is reached
3. Find the expected number of trees to sample from τ: $r\Phi = \sum_{i=1}^{k} r\phi_i$
4. Randomly round $r\Phi$
5. For each sample required:
   (a) Randomly choose an interval, i, according to the weights ϕ_i
   (b) Sample a tree uniformly at random from $S_i(\tau)$
6. Repeat from Step 2 until the required number of samples has been obtained.

Most n species trees based on real data will be a subsample of the m species contained in the true underlying tree such that $m - n$ species are missing. This problem is referred to as incomplete taxon sampling (Zwickl and Hillis 2002) and may be due to several reasons including inability to sample the species or a species being "undiscovered." If the number of missing species in a tree is substantial, incomplete taxon sampling should be included explicitly. A common approach is to sample trees with m species and randomly remove $m - n$ species, thus producing an n species tree as desired. For example, if only 75% of species are being sampled and we wish to sample a tree with 30 species, we would generate a tree with 40 species and remove 10 species uniformly at random. The problem with this approach is that we will generally only have an estimate of the number of missing species (25% in our example), hence we should consider a range of possible missing numbers of species. For instance, in the previous example, the true tree may have somewhere between, say, 35 and 50 species. GSA can readily be extended to include incomplete taxon sampling as discussed in Appendix 1.

Constant-Rate Birth–Death Approach

We have presented 2 main sampling approaches—PBMSA and GSA. PBMSA applies only to a limited class of evolutionary models that includes the Yule model and the coalescent model (for which PBMSA becomes equivalent to SSA). GSA applies to a much wider class of models including some for which SSA has been used inappropriately. Application of GSA to a given model is relatively straightforward regardless of the model's complexity. However, the generality of this approach makes it a mathematically unsatisfying and relatively inefficient process (from a computational perspective).

For the constant-rate birth–death process—an extension of the Yule model that explicitly incorporates extinction—the joint probability density of the speciation times can be inferred analytically (Yang and Rannala 1997). When the joint density of speciation events for a given model is known, a Markov chain Monte Carlo approach can be used to sample trees from this density.

However, in further work on the constant-rate birth–death process, we inferred the probability density for the time of individual speciation events explicitly (Gernhard 2008). In Appendix 2, we show how these densities can be used to sample trees from this model with the constant-rate BDSA—this is the most efficient way to sample trees from a constant-rate birth–death model of which we are aware. The BDSA for simulating trees on n species only requires to sample n times from one-dimensional distributions.

Comparison of the Sampling Approaches

We have shown that SSA is only appropriate for models without extinction where the time between speciation events is exponentially distributed with a rate parameter that depends only on the number of species that are extant. The 2 most popular models—the Yule and coalescent—satisfy these conditions, and it is appropriate to sample from them using SSA.

Existing approaches (such as SSA) are conceptually and computationally simpler than those introduced in this paper, they have also been applied to many situations in existing studies for which they are inappropriate. It is therefore important to consider how significantly the samples produced by the approaches differ. In situations where the difference is minimal, it may be appropriate to use the simpler existing approaches to produce an approximate sample. If the difference is great, it will be necessary to use more complicated approaches such as those presented here.

In this section, we investigate the differences resulting from the different simulation approaches. Note that throughout this section, we have disregarded the root edge length. We therefore define the age of a sampled tree as the distance between the speciation event that created the second species and the leaves. This corresponds to realistic situations where there it is often difficult to determine the length of the root edge.

Constant-Rate Birth–Death Model

We begin by comparing SSA and GSA using a constant-rate birth–death model—a simple extension

of the Yule model that explicitly includes extinction—each species has the same probability per unit time of becoming extinct. The constant-rate birth–death model includes 2 parameters—the speciation rate and the extinction rate—for our analysis, it is sufficient to consider the ratio of these, hence we set the speciation rate to 1. If the extinction rate is 0, the model is equivalent to the Yule model. By increasing the extinction rate from 0 to 1, the model becomes increasingly different from the Yule model and SSA should become increasingly inappropriate.

Figure 2 shows the expected age of the tree as a function of the extinction rate for samples of 10,000 trees produced by both sampling algorithms. When the extinction rate is 0, the model is equivalent to the Yule model and provides the same sample of speciation times. As the extinction rate increases, the age of the trees increases for 2 reasons: First, increasing the extinction rates effectively reduces the net speciation rate (speciation rate minus extinction rate), resulting in older trees—this effect is correctly incorporated by SSA as well as GSA. Second, tree age increases with a nonzero extinction rate, as we may return to n species from any tree on more than n species. The SSA should not be applied to models with a nonzero extinction rate because SSA only considers the first period when n species are present. The probability of returning from a larger number of species back to n species increases with increasing extinction rate; in fact, this probability is 1 in the extreme case, where the speciation and extinction rates are equal (as it is certain that the tree will eventually become extinct under that model). Figure 2 shows that the problem of returning multiple times to a tree on n species becomes severe for an extinction rate of more than 0.6 times the speciation rate.

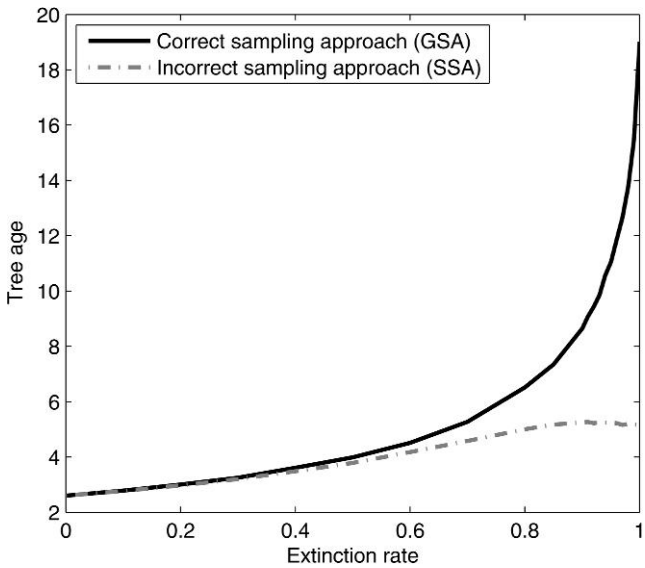

FIGURE 2. This figure shows the expected age for 20 species trees sampled from the constant-rate birth–death model as a function of the extinction rate. The speciation rate was set to 1, and 5000 trees were sampled for each extinction rate using SSA (dotted line) and GSA (solid line). The age of the trees sampled by GSA increases as the extinction rate increases—this is because SSA only considers the first time period during which n species existed, hence trees sampled using SSA do not exhibit the same age increase.

The correct simulation of the absolute times in a tree becomes important in studies where the absolute rates (e.g., millions of years) can be estimated for the data. Such an estimation is possible when fossil data are available (see, e.g., the paleontological studies Stanley et al. 1981; Raup 1991; Patzkowsky 1995; Przeworski and Wall 1998).

We have shown that the absolute age of the tree differs for the 2 sampling approaches; however, in some situations, the relative timing of the speciation events may be all that matters. To investigate this feature, we consider lineage through time (LTT) plots which show the number of species present as a function of the age of the tree. When the number of species is log transformed, the LTT plot should show a straight line with a deviation near the present (Harvey et al. 1994; Nee et al. 1994). Figure 3a shows the expectation of the LTT plot for an extinction rate of 0.95 from a sample of 10,000 trees produced using the 2 algorithms. There is a clear difference between GSA and SSA.

The slope near the origin (i.e., on the left) of a log-transformed LTT plot can be used to give an estimate of the net speciation rate. In Figure 3b, we consider the difference between this slope for the 2 methods as a function of the extinction rate. Interestingly, around an extinction rate of 0.9, the bias switches from negative to positive. This can be explained in the following way.

Under SSA, simulations stop at the first time when n species are reached. Because simulations under GSA can reach n species multiple times, simulations run longer and trees are older. When extinction rate is low, only recent speciation events disappear, but major clades survive. So the LTT plot shows no difference for SSA and GSA at the beginning, but toward the present, the GSA plot becomes flatter compared with the SSA plot. Therefore, when time is normalized, the slope near the origin is steeper under GSA than under SSA. As extinction rates increase, the number of older clades that go extinct increases due to much longer simulation run times, so the slope near the origin becomes smaller and smaller when using GSA as opposed to SSA, which eventually yields a change in slope bias.

Average extinction rates are generally found close to speciation rates (on the order of 0.9 or more, see, e.g., Alroy 1996, 2008), and this is a common setting for tree simulations (see, e.g., Magallón and Sanderson 2001; Ricklefs 2003). At this value, the 2 sampling approaches differ significantly in the estimated age of the tree. For the relative timing of speciation events, the result is not as clear, the severity (and direction) of the bias depends strongly on the extinction rate.

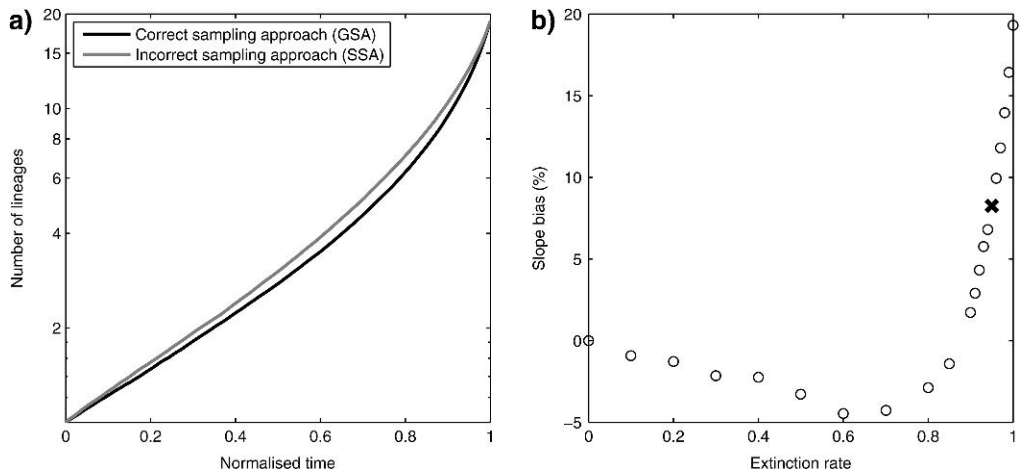

FIGURE 3. a) An expected LTT plot is shown here for 5000, 20-species trees sampled from a constant-rate birth–death model using both SSA and GSA. The speciation rate was set to 1 and the extinction rate to 0.95. The trees have been rescaled to have age 1—this removes the effect seen in Figure 2 and permits us to explore the relative speciation times of both samples. b) The initial slope in (a) gives an estimate of the net speciation rate. Here, we depict the percentage deviation of the slope obtained using SSA to that obtained with GSA for differing relative extinction rates. The point corresponding to (a) is marked.

Tree Shapes

The shape or topology of a tree is the structure obtained by disregarding the timing of speciation events (or equivalently the edge lengths). All memoryless models (including the constant-rate birth–death model) produce trees with the same tree shape distribution. The reason for this is that there is nothing to differentiate between species, hence, regardless of the model, each species is always equally likely to be the one that undergoes the next speciation or extinction event. Furthermore, because SSA does not distinguish between species, it correctly samples the tree shape distribution for memoryless models.

SSA may incorrectly sample the tree shape distribution from models that feature a memory. For pure birth models, the mechanism behind this would require a correlation between the shape of a tree and the duration for which n species exist. This correlation is not explicit in any common models of which we are aware but may exist implicitly; the strength of the correlation will determine the suitability of SSA to sample from a given model. We investigated 2 of the more common models with a memory (Heard 1996; Blum and Francois 2006) and found minimal bias in the tree shape distribution produced by SSA.

For other models, SSA may introduce a more serious bias in the tree shape distribution. One of the most obvious cases is a model with extinction where the tree shape distribution changes over time—as we have seen SSA produces trees that are too young, hence, the tree shape distribution would be sampled too early.

Incomplete Taxon Sampling

Let n be the number of sampled species in a tree. The most common approach for incomplete taxon sampling first samples a tree containing the expected true number of species, m, and then randomly deletes $m - n$ of these species. In Appendix 1, we provide an extension to GSA that considers a range of possible true tree sizes and samples these accordingly. We applied this method to the constant-rate birth–death model and found that the sampled trees differed negligibly from those obtained using the conventional approach. There are 2 main issues with the conventional approach; in this section, we illustrate why each issue results in only a negligible bias.

Issue 1: Consider the constant-rate birth–death model. Figure 4 shows how the expected age of a 10-species tree suffering from incomplete taxon sampling increases as a function of the true tree size. It is important to note that this is near-linear; in Stadler (2008), it is shown that for the constant-rate birth–death model the relationship is linear when the extinction rate is 1 and becomes slightly nonlinear as the extinction rate is decreased. If this relationship were perfectly linear and the true number of species were known, sampling a tree with m species and deleting $m - n$ species would give a correct sample. For this model, the deviation from linearity seems sufficiently small to be irrelevant for most purposes.

Issue 2: Given a probability of sampling each species (s), a naive method for calculating the expected number of species would be $m = n/s$. In Figure 4, we show the distribution of the true tree size as calculated using Equation A.4 for $s = 0.7$, due to the asymmetry of this distribution, its expectation exceeds n/s. In this example, the difference between these expectations is about 0.5; this will result in a small bias toward younger trees.

For the constant-rate birth–death model, the bias introduced by using a simplistic incomplete sampling method is insignificant in contrast with uncertainty regarding the true number of species. For other models, it

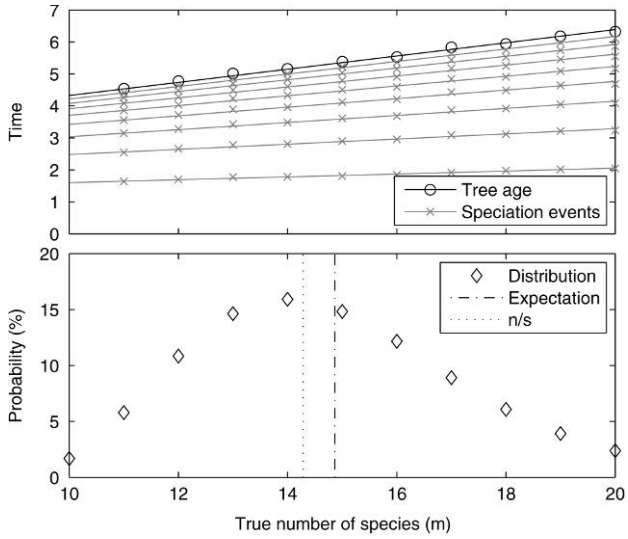

FIGURE 4. Top panel: The black circles show the expected age of a 10 -species tree that has been sampled by constructing an m species tree and deleting $m-10$ species. Five thousand samples were taken for each value of m using the constant-rate birth–death model with speciation rate 1 and extinction rate 0.9. The gray crosses show the expected time of the speciation events in the same situation. The lines are linear least squares fits to these points, demonstrating that the relationships are near-linear. The bottom panel shows the probability distribution of the true tree size, m, as calculated from Equation A.4 for a sampling probability of $s = 0.7$. Also depicted are the expectation of this distribution (about 14.8) and a simple estimate of this—n/s (about 14.3).

may be necessary to use the approach outlined in Appendix 1. This will particularly be the case for models that exhibit a strong nonlinearity in the expected age curve shown in Figure 4.

REANALYSIS OF PUBLISHED STUDIES WITH THE GSA MODEL

Lastly, we investigated if and how the use of inappropriate sampling methods influences conclusions in data analysis. We redid the analysis in Purvis (2004), a response paper to Ricklefs (2004). The 2 papers discuss if tree age and/or the number of species correlates with the variance of a trait in extant species. Two models of trait evolution are considered: gradual and punctuational evolution. Gradual trait evolution means that a trait changes continuously over time according to a Brownian motion. Punctuational trait evolution means that a trait only changes at a speciation event with the changes being normally distributed.

Ricklefs claimed that under the gradual trait evolution model, trait variance correlates with tree age, whereas under the punctuational trait evolution model, trait variance correlates with the logarithm of species number. Using this hypothesis, Ricklefs found that morphological evolution in birds is punctuational.

Purvis argued this result by simulating trees under the birth–death process: 1) with a fixed age and 2) with a fixed number of species. He then evolved the trait under both the gradual and the punctuational model and looked for correlations with the Pearson test.

Using the simulation results, Purvis showed for case (1) a strong correlation between the logarithm of species number and trait variance, under both the gradual and the punctuational trait evolution model. The simulation of birth–death trees under case (1) is straightforward: a tree is simulated until a certain age is reached. We will therefore discuss only case (2) further. For case (2), Purvis found a weak correlation between the tree age and trait variance, under both the gradual and the punctuational trait evolution model. He evolved 100 trees on 50 extant species using SSA (speciation rate 0.20, extinction rate 0.16). Then a Pearson test was performed in order to determine any correlation between the trait variance of the extant species and tree age. Under gradual trait evolution, he obtained $t_{98} = 2.03, P = 0.046$, and $r^2 = 0.03$ and under punctuational trait evolution $t_{98} = 2.32, P = 0.02$, and $r^2 = 0.04$. As the model had a nonzero extinction rate, the trees should have been sampled using an approach such as GSA instead of SSA. We reanalyzed these data using the GSA model to see whether the correlations inferred by Purvis were confirmed when using a more appropriate model.

First, we realized that the trait evolution simulations have a considerable variance. We therefore simulated 100 trees using SSA and GSA and then did Purvis's trait evolution analysis 20 times, that is, we simulated the trait 20 times for each tree. We then calculated the mean and standard deviation of the 20 values t_{98}, P, and r^2.

Using SSA for simulating the trees, under gradual trait evolution, we obtain $t_{98} = 1.92 \pm 0.94, P = 0.16 \pm 0.21$, and $r^2 = 0.04 \pm 0.03$ and under punctuational trait evolution $t_{98} = 1.47 \pm 0.98, P = 0.20 \pm 0.19$, and $r^2 = 0.03 \pm 0.03$. For the individual values, see Table 1.

Then, using our GSA for simulating the trees, under gradual trait evolution, we obtain $t_{98} = 3.07 \pm 0.94, P = 0.02 \pm 0.03$, and $r^2 = 0.09 \pm 0.05$ and under punctuational trait evolution $t_{98} = 2.32 \pm 1.45, P = 0.19 \pm 0.33$, and $r^2 = 0.07 \pm 0.05$. For the individual values, see Table 2.

There are 2 major conclusions from these simulations. First, there is high variance in the results of trait simulations on any one tree, as one would expect for these stochastic processes: some of our SSA runs are very similar to Purvis's, but the average results are not, see Table 1. Second, we found no overall correlation under SSA between tree age and trait variance under either the gradual or the punctuational trait evolution model. However, using the correct GSA, we find no correlation between tree age and trait variance under the punctuational trait evolution model, but we do find a weak correlation under the gradual trait evolution model. Note that Ricklefs (2004) suggested that such a correlation should only hold under the punctuational trait evolution model, whereas Purvis (2004) reported a correlation for both (implying that the data cannot be used to test for distinguish these models of trait evolution). Taken together, this suggests that more careful

TABLE 1. Using the SSA trees, we simulated 20 times a gradual evolving trait (g) as well as a punctual evolving trait (p)

Simulation	t_{98}, g	P, g	r^2, g	t_{98}, p	P, p	r^2, p
1	3.9309	0.0002	0.1362	0.9635	0.3377	0.0094
2	2.1835	0.0314	0.0464	3.8673	0.0002	0.1324
3	0.4033	0.6876	0.0017	1.6738	0.0974	0.0278
4	2.3020	0.0235	0.0513	1.5643	0.1210	0.0244
5	1.2827	0.2026	0.0165	1.6476	0.1026	0.0270
6	2.6989	0.0082	0.0692	2.6249	0.0101	0.0657
7	2.9981	0.0034	0.0840	0.9679	0.3355	0.0095
8	1.2641	0.2092	0.0160	2.0306	0.0450	0.0404
9	0.9015	0.3695	0.0082	0.7662	0.4454	0.0060
10	1.0374	0.3021	0.0109	1.7590	0.0817	0.0306
11	2.1176	0.0367	0.0438	−0.9682	0.3353	0.0095
12	2.1256	0.0361	0.0441	0.7035	0.4834	0.0050
13	3.0446	0.0030	0.0864	1.4626	0.1468	0.0214
14	1.5038	0.1359	0.0226	2.4407	0.0165	0.0573
15	0.4809	0.6316	0.0024	2.2495	0.0267	0.0491
16	2.7056	0.0080	0.0695	1.6299	0.1063	0.0264
17	2.6804	0.0086	0.0683	0.5717	0.5689	0.0033
18	2.0277	0.0453	0.0403	0.6506	0.5168	0.0043
19	1.0014	0.3191	0.0101	1.3566	0.1780	0.0184
20	1.7940	0.0759	0.0318	1.5209	0.1315	0.0231
Mean	1.9242	0.1569	0.0430	1.4741	0.2043	0.0295
Standard deviation	0.9358	0.2083	0.0348	0.9778	0.1854	0.0300

TABLE 2. Using the GSA trees, we simulated 20 times a gradual evolving trait (g) as well as a punctual evolving trait (p)

Simulation	t_{98}, g	P, g	r^2, g	t_{98}, p	P, p	r^2, p
1	3.5009	0.0007	0.1112	0.6701	0.5044	0.0046
2	4.3062	0.0000	0.1591	3.3604	0.0011	0.1033
3	2.6522	0.0093	0.0670	2.6732	0.0088	0.0680
4	4.5924	0.0000	0.1771	2.4367	0.0166	0.0571
5	2.2197	0.0288	0.0479	4.1838	0.0001	0.1515
6	3.5131	0.0007	0.1119	2.3722	0.0196	0.0543
7	3.5539	0.0006	0.1142	0.4154	0.6788	0.0018
8	2.2403	0.0273	0.0487	3.1152	0.0024	0.0901
9	4.8480	0.0000	0.1934	4.9537	0.0000	0.2003
10	3.4050	0.0010	0.1058	3.0170	0.0033	0.0850
11	3.0104	0.0033	0.0846	2.5407	0.0126	0.0618
12	1.5544	0.1233	0.0241	3.3926	0.0010	0.1051
13	1.9690	0.0518	0.0381	−0.0532	0.9577	0.0000
14	3.7756	0.0003	0.1270	4.1254	0.0001	0.1480
15	2.6184	0.0102	0.0654	2.5801	0.0114	0.0636
16	4.0335	0.0001	0.1424	2.3012	0.0235	0.0513
17	2.0592	0.0421	0.0415	2.2018	0.0300	0.0471
18	2.4036	0.0181	0.0557	−0.2872	0.7746	0.0008
19	2.0635	0.0417	0.0416	2.0397	0.0441	0.0407
20	2.9868	0.0036	0.0834	0.3439	0.7316	0.0012
Mean	3.0653	0.0181	0.0920	2.3191	0.1911	0.0668
Standard deviation	0.9445	0.0299	0.0494	1.4520	0.3278	0.0553

simulations using the proper GSA will be required to discern if and how one can use clade age, size, and trait variance to distinguish between gradual and punctuational trait evolution.

CONCLUDING COMMENTS

When exploring evolutionary models, analytical results are preferable to simulations because of the reduced computational burden and greater insight they provide. However, analytical results may be difficult to obtain and simulation studies may answer questions more quickly—once a result has been confirmed by simulation studies, an analytical approach can be pursued with greater confidence.

Simulation studies have an inherent danger—it is extremely easy to simulate trees using a given model, however understanding what distribution these trees come from can be difficult. This makes it easy to proceed with a (possibly incorrect) method and therefore sample trees. This is particularly problematic with more complicated evolutionary models where seemingly intuitive methods of simulating trees (such as SSA) often sample from undesirable and unrealistic probability distributions.

We have shown that a commonly used sampling approach (SSA) is appropriate for 2 of the most common evolutionary models—the Yule model and coalescent model. However, this approach is inappropriate for many other models to which it has been applied. We developed an appropriate GSA and considered the biases introduced when SSA is used instead of an appropriate approach such as GSA. For the constant-rate birth–death model, SSA produces a strong bias in the age of the tree and the relative timing of speciation events. It does not produce a bias in the tree shape distribution. Furthermore, for the constant-rate birth–death model, using GSA with the common approach for incorporating incomplete taxon sampling seems adequate for most applications. More complex models with certain characteristics as discussed in this paper may result in stronger biases of any of these attributes of a sampled tree.

Finally, we have shown that using SSA rather than GSA leads to qualitatively different conclusions: By simulating traits under a gradual trait evolution model on simulated birth–death trees, we find no correlation between trait variance and tree age using SSA, whereas we find a weak correlation using GSA. We suggest that many analyses based on SSA may need to be reanalyzed with the more appropriate GSA.

The methods presented here have been implemented in a Perl package, including a user friendly GUI (Hartmann 2010), and in an R package, TreeSim (Stadler 2010). This software has built-in support for the Yule model and constant-rate birth–death models and is easily extendable to permit sampling from additional models. We hope that this paper helps clarify some of the issues about sampling trees from evolutionary models and that the software we have created will be of use for future simulation studies.

FUNDING

The authors wish to thank the The Allan Wilson Centre for Molecular Ecology and Evolution for facilitating this collaborative work through travel grants and scholarships.

ACKNOWLEDGEMENTS

We wish to thank Kristine Hartmann, Todd Oakley, Arne Mooers, Mike Steel, Jack Sullivan, Tim Collins, and the two anonymous referees for their comments on this manuscript.

REFERENCES

Agapow P.-M., Bininda-Edmonds O.R., Crandall K.A., Gittleman J.L., Mace G.M., Marshall J.C., Purvis A. 2004. The impact of species concept on biodiversity studies. Q. Rev. Biol. 79:161–179.

Aldous D., Popovic L. 2005. A critical branching process model for biodiversity. Adv. Appl. Probab. 37:1094–1115.

Alroy J. 1996. Constant extinction, constrained diversification, and uncoordinated stasis in North American mammals. Palaeogeogr. Palaeoclimatol. Palaeoecol. 127:285–311.

Alroy J. 2008. Dynamics of origination and extinction in the marine fossil record. Proc. Natl. Acad. Sci. USA. 105:11536.

Bininda-Edmonds O.R.P., Cardillo M., Jones K.E., MacPhee R.D.E., Beck R.M.D., Grenyer R., Price S.A., Vos R.A., Gittleman J.L., Purvis A. 2007. The delayed rise of present-day mammals. Nature. 446:507–512.

Blum M., Francois O. 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst. Biol. 55:685.

Chan K.M.A., Moore B.R. 1999. Accounting for mode of speciation increases power and realism of tests of phylogenetic asymmetry. Am. Nat. 153:332–346.

Chan K.M.A., Moore B.R. 2002. Whole-tree methods for detecting differential diversification rates. Syst. Biol. 51:855–865.

Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–778.

Hahn M., De Bie T., Stajich J., Nguyen C., Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 15:1153.

Harding E.F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Probab. 3:44–77.

Hartmann K. 2010. Tree sample. Hobart, Australia: Tasmanian Aquaculture and Fisheries Institute, University of Tasmania. Available from: http://treesample.googlecode.com.

Harvey P.H., May R.M., Nee S. 1994. Phylogenies without fossils. Evolution. 48:523–529.

Heard S.B. 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rates. Evolution. 50:2141–2148.

Hey J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution. 46:627–640.

Hohl M., Ragan M.A. 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst. Biol. 56:206–221.

Karev G.P., Wolf Y.I., Koonin E.V. 2003. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? Bioinformatics. 19:1889–1900.

Kingman J.F.C. 1982a. Exchangeability and the evolution of large populations. In: Koch G., Spizzichino F., editors. Exchangeability in probability and statistics. Amsterdam-New York: North-Holland. p. 97–112.

Kingman J.F.C. 1982b. On the genealogy of large populations. J. Appl. Probab. 19A:27–43.

Kingman J.F.C. 1982c. The coalescent. Stoch. Processes Appl. 13: 235–248.

Magallón S., Sanderson M. 2001. Absolute diversification rates in angiosperm clades. Evolution. 55:1762–1780.

Mooers A.Ø., Harmon L., Blum M., Wong D., Heard S. 2007. Some models of phylogenetic tree shape. In: Gascuel O., Steel M., editors. Reconstructing evolution—new mathematical and computational advances. Oxford, UK: Oxford University Press.

Mooers A.Ø., Heard S.B. 1997. Inferring evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72:31–54.

Moran P. 1958. A general theory of the distribution of gene frequencies. I. Overlapping generations. Proc. R. Soc. Lond. B. 149:102–112.

Nee S., Holmes E., May R., Harvey P. 1994. Extinction rates can be estimated from molecular phylogenies. Philos. Trans. R. Soc. Lond. Biol. Sci. 344:77–82.

Oakley T., Ostman B., Wilson A. 2006. Repression and loss of gene expression outpaces activation and gain in recently duplicated fly genes. Proc. Natl. Acad. Sci. USA. 103:11637.

Patzkowsky M. 1995. A hierarchical branching model of evolutionary radiations. Paleobiology. 21:440–460.

Popovic L. 2004. Asymptotic genealogy of a critical branching process. Ann. Appl. Probab. 14:2120–2148.

Przeworski M., Wall J. 1998. An evaluation of a hierarchical branching process as a model for species diversification. Paleobiology. 24: 498–511.

Purvis A. 2004. Evolution: how do characters evolve? Nature. 432:7014.

Pybus O., Harvey P. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. B. 267: 2267–2272.

Rambaut A. 2002. PhyloGen: phylogenetic tree simulator package. Department of Zoology, University of Oxford, Oxford, UK. Available from: http://tree.bio.ed.ac.uk/software/phylogen/.

Raup D. 1991. A kill curve for phanerozoic marine species. Paleobiology. 17:37–48.

Ricklefs R. 2003. Global diversification rates of passerine birds. Proc. R. Soc. Lond. B. 270:2285–2291.

Ricklefs R. 2004. Cladogenesis and morphological diversification in passerine birds. Nature. 430:338–341.

Sepkoski J.J. 1982. Mass extinctions in the Phanerozoic oceans: a review. Geol. Soc. Am. Spec. Pap. 190:281–289.

Shaw A., Cox C., Goffinet B., Buck W., Boles S. 2003. Phylogenetic evidence of a rapid radiation of pleurocarpous mosses (bryophyta). Evolution. 57:2226–2241.

Stadler T. 2008. Lineages-through-time plots of neutral models for speciation. Math. Biosci. 216:163–171.

Stadler T. 2010. Treesim. Institute of Integrative Biology, Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland. Available from: http://www.tb.ethz.ch/people/tstadler.

Stanley S.M., Signor P.W. III, Lidgard S., Karr A. F. 1981. Natural clades differ from "random" clades: simulations and analyses. Paleobiology. 7:115–127.

Tanaka M., Francis A., Luciani F., Sisson S. 2006. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics. 173:1511.

Venditti C., Meade A., Pagel M. 2006. Detecting the node-density artifact in phylogeny reconstruction. Syst. Biol. 55:637–643.

Weir J.T. 2006. Divergent timing and patterns of species accumulation in lowland and highland neotropical birds. Evolution. 60:842–855.

Yang Z., Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. 17:717–724.

Yule G.U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos. Trans. R. Soc. Lond. B. 213:21–87.

Zwickl D.J., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

APPENDIX 1: EXTENSION OF GSA TO INCOMPLETE TAXON SAMPLING

Here, we extend GSA to explicitly take into account incomplete taxon sampling. This extension of GSA requires either an estimate of the probability, s, of any given species being sampled or, alternatively, the probability distribution of the size of the true tree, m, given the number of sampled species, n. Without one of these quantities, our method cannot be applied, and indeed, it is difficult to see how to proceed otherwise. Our method also assumes that sampled species are uniformly at random distributed through the tree. It is relatively straightforward to relax this last assumption although we do not present the details here. One instance where this would be necessary is if the probability of sampling any 2 species is positively correlated to their proximity in the phylogenetic tree (as might be the case if whole clades are likely to be missed or thoroughly sampled).

Given the sampling probability s, for a given real tree size, m, the number of sampled species, n, will be

distributed according to a binomial distribution:

$$p(n|m) = \binom{m}{n} s^n (1-s)^{m-n}. \quad \text{(A.1)}$$

However, the number of sampled species, n, is the size of the final tree and is what we wish to condition on; thus, Bayes' law gives us

$$p(m|n) \propto p(n|m)p(m), \quad \text{(A.2)}$$

where $p(m)$ is the probability of a tree having m leaves and $p(n|m)$ is the probability of sampling n of those leaves. For $m \geq n$, it is always possible to obtain n leaves from a tree with m leaves, however the probability of this occurring decreases with m such that $p(n|m)$ becomes small enough to make $p(m|n)$ negligible. This permits us to restrict the range of m that must be examined to $n \leq m \leq m^*$, where m^* is a limit that needs to be established. If we assume that $p(m)$ does not increase with m, an appropriate condition to solve for m^* is

$$p(n|m^*) \leq \sum_{m=n}^{m^*-1} \frac{p(n|m)}{N}, \quad \text{(A.3)}$$

where N is the number of trees that are being sampled. This condition ensures that the first value of m being excluded is expected to contribute less than 1 tree to the final sample. If $p(m)$ increases with m, extra analysis will be required to find an appropriate m^* (e.g., using simulation studies).

Given a particular simulated tree, we have $p(m) \propto \Phi_m$ (the duration for which a simulated tree had m species), hence substitution in Equation A.2 gives

$$p(m|n) \propto \Phi_m \binom{m}{n} s^n (1-s)^{m-n}, \quad \text{(A.4)}$$

which is readily normalized to give $p(m|n)$. The expected contribution to the sample from a given simulated tree consists of the expected contribution for each value of m:

$$r \sum_{m=n}^{m^*} \Phi_m p(m|n). \quad \text{(A.5)}$$

When a tree is simulated, the expected contribution to the sample is found and a sample of the corresponding size is taken. This process is repeated until the sample has the desired size.

GSA with incomplete taxon sampling

1. Find m^* analytically or by simulation/investigation (e.g., Equation A.3)
2. Simulate a tree τ until m^* species are reached or all species become extinct
3. Calculate $p(m|n)$ for all m for this simulated tree (Equation A.4)
4. Find the expected number of samples to take from τ (Equation A.5)
5. Randomly round the expected number of samples
6. For each sample:
   (a) Randomly choose the original tree size $\hat{m}$ according to $p(m|n)$
   (b) Uniformly at random choose a time when τ had $\hat{m}$ species
   (c) Randomly delete $\hat{m} - n$ species
7. Repeat from Step 2 until all samples have been obtained.

Appendix 2: Efficient Sampling from the Constant-Rate Birth–Death Model

In this section, we present an efficient algorithm for sampling trees with n species from the constant-rate birth–death model. The constant-rate birth–death model is a popular null model for detecting variation in diversification rates (Mooers and Heard 1997; Pybus and Harvey 2000; Chan and Moore 2002). It is an extension of the Yule model where all species have a constant rate of speciation, β, and a constant rate of extinction, μ, with the constraint that $\beta \geq \mu$.

The method we propose relies on representing a binary tree as a point process, this is illustrated in Figure 5. Generally, a binary tree with n extant species can be described by $n-1$ points in the following way. On a horizontal axis, locate the leaves (species) uniformly at random at $1, 2, \ldots n$. The $n-1$ speciation times are represented by $n-1$ points with (x, y) coordinates $(j+1/2, s_j)$, $j = 1, 2, \ldots, n$; $s_j > 0$. The tree is obtained by an iterative procedure. At each step of the iteration, the most recent speciation event is connected with the 2 neighboring leaves. This speciation event is regarded as a new leaf and replaces the 2 neighboring leaves. This is repeated until all speciation points are connected.

In Gernhard (2008), it is shown that the times s_i of the speciation events in a constructed tree under the constant-rate birth–death model are independent and identically distributed. For $\beta > \mu$, we have the distribution function

$$F(s|t, \beta, \mu, n) = \frac{1 - e^{-(\beta-\mu)s}}{\beta - \mu e^{-(\beta-\mu)s}} \frac{\beta - \mu e^{-(\beta-\mu)t}}{1 - e^{-(\beta-\mu)t}},$$

where t is the time of origin of the tree. The inverse of $F(s|t, \beta, \mu, n)$ is

$$F^{-1}(s|t, \beta, \mu, n) = \frac{1}{\beta - \mu} \ln \left(\frac{\beta - \mu e^{-(\beta-\mu)t} - \mu(1 - e^{-(\beta-\mu)t})s}{\beta - \mu e^{-(\beta-\mu)t} - \beta(1 - e^{-(\beta-\mu)t})s} \right).$$

Recall that throughout this paper, we assume a uniform prior for the time of origin of a tree. For this approach, we need the probability density of the time of origin of the tree, t, conditional on it having n species at the present. This distribution was derived in Gernhard (2008) for $\beta > \mu$:

$$Q(t|\beta, \mu, n) = \left(\frac{\beta(1 - e^{-(\beta-\mu)t})}{\beta - \mu e^{-(\beta-\mu)t}} \right)^n.$$

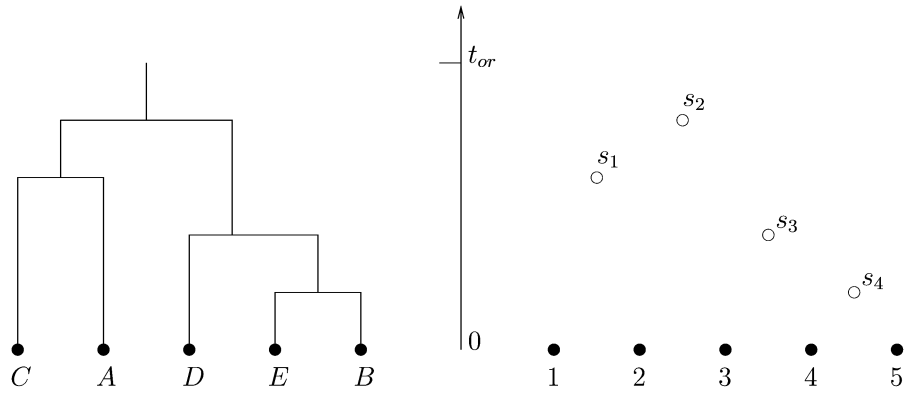

FIGURE 5. On the left, a tree on 5 species—A, B, C, D, and E—is displayed. Time is set 0 at the time of the leaves and increasing into the past. The time t_{or} is the origin of the tree. On the right, we have the corresponding point process representation. For simulating trees under the constant-rate birth–death process, first the times s_i and t_{or} are sampled. Then the point process is transformed into a tree as described in Appendix 2, and the species are assigned uniformly at random to $\{1, 2, \ldots, n\}$.

The inverse of Q is

$$Q^{-1}(t|\beta, \mu, n) = \frac{1}{\beta - \mu} \ln \left(\frac{1 - \frac{\mu}{\beta} t^{1/n}}{1 - t^{1/n}} \right).$$

For $\beta = \mu$, the functions $F(s|t, \beta, \beta, n)$ and $Q(t|\beta, \beta, n)$ are established in Aldous and Popovic (2005):

$$F(s|t, \beta, \beta, n) = \frac{s}{1 + \beta s} \frac{1 + \beta t}{t},$$

$$F^{-1}(s|t, \beta, \beta, n) = \frac{st}{1 + \beta t(1 - s)},$$

$$Q(t|\beta, \beta, n) = \left(\frac{\beta t}{1 + \beta t} \right)^n,$$

$$Q^{-1}(t|\beta, \beta, n) = \frac{1}{\beta(t^{-1/n} - 1)}.$$

Combining these probability densities and the point process representation, we obtain the following algorithm:

Constant-rate BDSA

1. Sample $r_0, \ldots, r_{n-1}$ uniformly at random from $[0, 1]$
2. Calculate the age of the tree, $t = Q^{-1}(r_0|\beta, \mu, n)$
3. Calculate the $n-1$ branching times, $s_i = F^{-1}(r_i|t, \beta, \mu, n), i = 1, \ldots, n-1$
4. Construct the tree from the point process representation
5. Repeat from Step 1 until all samples have been obtained.

The advantage of this method over GSA is that it is unnecessary to determine n^* and r. The disadvantage of this method is that it gives no information about extinct lineages (regardless of the value of μ). If this information is required, GSA must be used for sampling constant-rate birth–death models. Finally, note that a sample from the Yule model can be obtained by setting $\mu = 0$.